SOME RESEARCH PROBLEMS IN AUTOMATIC INFORMATION RETRIEVAL

G. Salton*
Department of Computer Science
Cornell University
Ithaca, New York 14853

Abstract

Information retrieval components are currently incorporated in several types of information systems, including bibliographic retrieval systems, data base management systems and question-answering systems. Some of the problems arising in the real-time environment in which these systems operate are briefly discussed. Certain recent advances in information retrieval research are then mentioned, including the formulation of new probabilistic retrieval models, and the development of automatic document analysis and Boolean query processing techniques.

1. Types of Information Systems

It is customary to distinguish three main types of information systems, known as data base management systems, bibliographic reference retrieval systems, and question-answering systems, respectively. On the surface the problems which arise in these three areas are very much the same, in the sense that in each case stored information files are processed in response to queries submitted by a population of users, and that answers are generated to these queries. In practice, the workers in these three areas belong to distinct communities and the problems they have chosen to work with are different.

<u>Data base management</u> is concerned with the processing of structured data files normally represented by two-dimensional tables and by relationship indications between different tables. Each row of such a table may be used to represent a record included in the file, each column identifies some attribute whose values are then used to distinguish the records from each other. Since the mathematical notion of an n-ary relation between sets is precisely equivalent to the collection of rows of a table, each consisting of the values

*This study was supported in part by the National Science Foundation under grant IST 80-17589.

© 1983 ACM 0-89791-107-5/83/006/0252 $00.75

of n particular attributes, the mathematical theory of relations can be applied to the manipulations of files in data base management; the corresponding equivalences has been used with great effect in recent research in the data base field.

The problems of particular interest to researchers in data base management deal with the efficient processing of structured data represented by tables [1,2]. A typical list of data base management problems appears in Table 1.

a) retrieval of specific data from a set of structured tables;

b) high-level descriptions and models of the structured data files and of their relationships;

c) high-level descriptions of processes designed to manipulate the structured data;

d) high-level descriptions of languages designed to formulate data base queries;

e) formulation of particular constraints between individual elements stored in structured files--for example, the fact that each person in an employee file has a particular age specified by a number between 15 and 65;

f) implementation of concurrent operations on structured data to simulate the situation where different users desire to obtain access to files simultaneously;

g) implementation of operations on distributed files where the individual components of a data base management system are stored in different locations possibly attached to different computers.

Typical Problems of Interest in Data Base Management

Table 1

The main distinction to be drawn between *bibliographic retrieval* and data base management is that in bibliographic retrieval the aim is to retrieve bibliographic references, that is, citations to bibliographic items, rather than actual data. This implies that for retrieval purposes it is not generally necessary to access the individual components of a given stored item, but only the general citation. However, in order to identify a particular citation it becomes necessary to specify the information content of the items. In data base management, the content analysis problem is by-passed by using a set of prespecified attributes such as age, name, or address for each record. However, the content analysis problem takes on major importance in bibliographic text processing.

Since many bibliographic retrieval files include hundreds of thousands, or even millions of items, a great deal of work in reference retrieval is devoted to the construction of index structures designed to provide rapid access to the individual items responding to particular search specification. A number of problems of current interest in bibliographic information retrieval are listed in Table 2. [3,4]

a) automatic indexing techniques designed to extract effective content identifiers from natural language texts;

b) automatic language normalization techniques designed to broaden the indexing vocabulary by means of thesauruses, or to narrow the vocabulary by phrase assignments;

c) automatic classification methods capable of assigning the items to affinity classes of similar items;

d) search methodologies designed to retrieve particular classes of items in response to incoming requests, including fast text scanning methods, searches using auxiliary indexes, and clustered searches which proceed through a hierarchial arrangement of record clusters;

e) automatic query formulation methods capable of generating useful arrangements of query terms, possibly interconnected by Boolean operators, and capable of retrieving relevant items and rejecting nonrelevant ones.

Typical Problems of Interest in Bibliographic Reference Retrieval

Table 2

Data base management systems manipulate simple files represented by two-dimensional tables; reference retrieval systems, on the other hand, process unstructured text files. For this reason it becomes necessary in bibliographic information retrieval to deal with automatic language processing methods. This is true to an even greater degree in automatic *question-answering* where the aim is to furnish direct answers to user queries often submitted in natural language formulations. In this sense, question-answering systems constitute a bridge between data base management and bibliographic retrieval, since actual data must be retrieved as in data base management, while operating in a natural language environment as in bibliographic retrieval.

In reference retrieval it is often sufficient to characterize each bibliographic item by using a few content terms reflecting the main subject areas of interest. In question-answering information queries dealing with very particular facts may have to be answered directly. In these circumstances it becomes necessary to use a much more detailed language analysis system. Some problem areas of interest in question-answering are listed in Table 3. [5-7]

a) the analysis of information requests using syntactic, semantic, and inferencing techniques to determine content and focus of a query;

b) the construction of knowledge-bases designed to store the facts of interest in a given subject area;

c) the storage of the general world knowledge which is assumed in a given question-answering situation--for example, the fact that pigs are large and writing pens are small compared to pig pens;

d) the generation of new facts derivable from other facts already known, and the processing of imperfectly specified information;

e) the search through complex information file storing heterogeneous items with multiple relationships between items.

Typical Problems of Interest in Automatic Question-Answering

Table 3

Because the question-answering problems are difficult to solve, operating question-answering systems have been implemented only in microworlds with restricted subject areas, and simplified language processing rules. Moreover, the methods used to handle the analysis in one subject area have not normally been extendable to other larger areas. In the recent past, attempts have been made to devise memory structures of some generality and to implement generally usable processing techniques. [6,7] There has also been some progress in relating bibliographic retrieval more closely with question-answering, notably in the work on *passage retrieval*, where direct answers to questions are obtainable by retrieving certain passages (sentences) of texts rather than complete texts as in reference retrieval. [8-10]

2. Some Problems in Information Network Processing

It was noted in the previous section that the various types of information processing systems are related by a common requirement to retrieve information in response to questions submitted by a user population. Another common feature is the physical environment in which most retrieval activities take place. In most operational situations, *on-line access* is provided to the stored data by using terminal equipment directly connected to the retrieval processor and to the information store. Furthermore, since many difference consoles can be connected to the information resources, a so-called *information network* results through which access to stored data bases can be shared by many users. [12,13] The construction of information networks servicing heterogeneous user populations raises many interesting questions concerning information security, the preservation of data integrity and correctness, and information privacy.

The basic processing environment consisting of individual user terminal equipment connected to computerized information files can be supplemented by *local microprocessor* devices permitting the users to store and process individual, private files locally, and thus to combine local processing conducting on private files with remote processing using public files. [14,15] Among the information services which may be provided in such an environment, the following may be of major importance: [16] current awareness and automatic reminding services, on-demand search services, message handling facilities, text search capabilities, file creation and text composition methods, word processing and text editing facilities, and text publication services.

A great deal has been written about the advantages of paperless information systems: the decreased cost and reduced storage size compared with paper products, the improved search capabilities, the decreased error rates resulting from the use of error detecting and correcting schemes, the automatically generated search aids in the form of dictionaries and indexes of various kinds, the access to scarce or expensive resources which would otherwise be inaccessible. However, many problems remain to be overcome in information system design, the most obvious of which deals with the mismatch between the requirements of the sophisticated automatic systems, and the relative ignorance of many users. Efforts have been made to build user-friendly interfaces to lessen the strain of using the automatic systems; but more remains to be done in helping people to access the available facilities. [17,18] A list of problem areas in information system design is included in Table 4.

3. Research Problems in Information Retrieval

A) Information Retrieval Models

A great deal of interesting work has been done in recent years in information system modelling. [19-22] Of particular interest in this connection are the *probabilistic models* whose introduction has led to the creation of optimal term weighting systems and to the use of associations and dependencies between terms. [23-27] In probabilistic information retrieval, the information search and retrieval problem is reduced to a probability extimation problem concerning the relevance or nonrelevance of the individual documents in a collection. Unfortunately, the relevance properties of individual documents cannot be determined in the abstract. Instead, it becomes necessary to relate the relevance probability of each document to the occurrences of the individual document terms assigned as content identifiers to the relevant and nonrelevant documents of a collection.

Probabilistic retrieval models have been constructed using the following principal assumptions concerning the occurrence characteristics of the terms in the documents of a collection:

a) The simplest model assumes that the terms all occur independently of each other in the relevant and nonrelevant documents of a collection. The term independence model has been used in practice with substantial success. This model can be used directly to rank the documents for retrieval purposes and also to derive optimum weighting functions for query terms, known as *term relevance* weights. [28-29] The term relevance weights may in turn be approximated in some circumstances by the well-known *inverse document frequency* weights which have proved particularly simple and useful in indexing and query formulations. [30-31]

b) The next model in order of complexity assumes that each term assigned to a document depends on at most one other term. This gives rise to the tree dependence model. [32-34] Since the tree dependence model takes into account relationships between certain term pairs in addition to the single term probabilities, one may expect that better retrieval results are obtainable with tree dependence than with the term independence model.

c) A still more refined system consists in using dependencies between certain term triples in addition to the single and pairwise probability factors. Such an extended tree dependence model has recently been used with some success. [35]

d) A complete probability model which takes into account the dependencies between all subsets of terms may be based on the use of the Bahadur Lazarsfeld expansion. [36]

a) the design of unified information systems capable of carrying out data base as well as bibliographic reference retrieval activities and of answering questions in specific subject areas;

b) the design of user-friendly interfaces making it possible for untrained users to interact with the automatic system without undue hardship;

c) the construction of flexible classification systems and multiple information accessing techniques to bridge the gap between the user's view of the information store and the system view;

d) the development of information comparison and text abstracting methods to reduce the size of the stored information files;

d) the institution of fast text scanning, and document skimming methods capable of rejecting rapidly the mass of extraneous materials;

f) the generation of restricted access electronic mail and message systems to enable each individual to restrict the type of incoming messages to only those which are actually wanted;

g) the assignment of appropriate roles to the publishing industry and to the conventional library organizations to insure that the special advantages of those organizations are maintained in future automatic information environments.

Typical Problems in Information System Design

Table 4

The probabilistic retrieval models are not usable in practice unless the occurrences characteristics of the individual terms in both the relevant and nonrelevant documents of a collection can be estimated with reasonably accuracy. One way of performing this estimation process is to use an iterative search strategy based on *relevance feedback* where some previously retrieved documents are identified by the users as relevant, or nonrelevant to their search request. In that case the term occurrence probabilities can be approximated by the actually observed occurrence probabilities in the previously retrieved documents identified as relevant and nonrelevant, respectively. [37]

Some of the probabilistic models described in the literature have recently been compared and unified [38], and a new, ultimate probabilistic model has been proposed which makes maximum use of all available information without implicitly making assumptions about any unknown data. [39] This last model appears to be computationally difficult, but further progress may be anticipated in the design and use of probabilistic retrieval models.

B) Advanced Boolean Retrieval Systems

The probabilistic indexing and retrieval models examined in the previous section can be used to obtain term weight for the terms assigned to a document or query as a function of the occurrence frequencies of the terms in the relevant and nonrelevant documents of a collection. The actual choice of index terms is not, however, provided or part as the model. For this purpose completely *automatic indexing* methods are available which are capable of assessing the term specificity of potential content terms, and of assigning either single terms of the correct specificity, or term phrases for combinations of high frequency terms, or thesaurus classes for groupings of low frequency terms. [40,41]

The automatic indexing methods serve for the assignment of content identifiers to the *documents* of a collection. The formulation of the search requests requires additional steps, notably the choice of Boolean operators (*and*, *or*, *not*) to relate the various query terms. The use of Boolean queries involves a number of disadvantages not the least of which is the difficulty of formulating good Boolean query statements. In addition, the conventional Boolean processing technology does not allow the use of weighted terms, and will not produce ranked document output. Finally the use of Boolean conventional logic may be counter-intuitive in information retrieval in some circumstances.

A good deal of work has been done in an attempt to generalize and improve the Boolean query processing strategies. For example, *fuzzy set models* have been proposed which are compatible with the standard Boolean logic and also make it possible to assign weights to the terms assigned to the documents (but not to the queries) of a collection. [42-45] The fuzzy set models unfortunately exhibit the same disadvantages as the ordinary Boolean processing models in the sense that retrieval of a document in response to an *or*-query depends on a single query term only, and retrieval in response to an *and*-query depends on the full set of query terms.

A more flexible, so-called *extended Boolean* query processing system has been proposed recently based on a relaxed system of interpretation for the Boolean operators. [46] In the extended system the Boolean operators can be treated less strictly than in a conventional system: in particular, the presence of more query terms in a given document is worth more than the presence of fewer query terms. Specifically, a generalized distance function based on Lp vector norms is used to compare the Boolean queries with the documents identified by sets of content terms. This distance function involves a parameter p, which can be made to vary from one to infinity with the following results:

a) When $p = \infty$, the Boolean operators are treated strictly, as in conventional Boolean logic.

b) As p decreases from infinity, the operators and and or become less and less strict. Thus an and-operator with a p value equal to 10 will favor the presence of most rather than all query terms in a document; an or-operator with a p-value equal to 10 will favor the presence of some query terms in a document, rather than the presence of one query term.

c) As p reaches its lower limiting value of 1, the and and or operators are relaxed to such an extent that the distinction between them is lost completely, and queries (A and B) and (A or B) are treated simply as the vector query (A,B).

The extended Boolean retrieval system exhibits the following advantages for a practical implementation:

a) it accommodates the normal query structure used in conventional Boolean retrieval but avoids potentially nonsensical interpretations by letting the query-document similarity depend on the number and the weight of the matching terms;

b) it allows the incorporation of term weights into both documents and queries in accordance with the presumed importance of the terms in the respective constructs;

c) it provides for the retrieval of documents in decreasing order of the query-document similarity, thereby providing control of the size of the document set to be retrieved in response to a given query;

d) it facilitates the distinction between compulsory phrase and strict synonym interpretations on the one hand, and tentative phrases and looser synonym relations on the other, by using high p-values to characterize the query structures in the former case, and low p-values in the latter;

e) it is compatible with conventional inverted file technologies by using completely automatic Boolean query construction methods based on the available natural language statements of user needs, or on the texts of previously retrieved documents identified as relevant.

A strategy for carrying out information searches in the extended Boolean query processing environment is outlined in Table 5.

a) construct a conventional Boolean query (or use an available Boolean query) which is broad enough to retrieve a large proportion of the potentially relevant documents;

b) process this query against the available document collection using a conventional inverted file system thereby retrieving a subset of documents D' included in the collection;

c) use the original natural language statement of user need, or the original conventional Boolean query obtained in step a) to construct a relaxed query in the extended Boolean system with low p-values ($1 \le p \le 3$);

d) process the extended Boolean query obtained in step c) against the collection D' obtained in step b) and rank the documents in decreasing query-document similarity order;

e) examine some items retrieved in step d) and use terms included in the retrieved items identified as relevant to construct an improved extended Boolean query; return to step d) and repeat until the user is satisfied with the output.

Basic Processing Steps in Extended Boolean Retrieval System

Table 5

A number of open problems and advances in information retrieval have been discussed in this note, including the design of unified retrieval environments for different information structures, the operations of on-line paperless information networks, the use of probabilistic retrieval models, and the design of user friendly systems for Boolean query processing. One may expect that substantial progress will be reported in all these areas in the foreseeable future.

References

[1] C.J. Date, An Introduction to Database Systems, Third Edition, Addison Wesley Publishing Company, Reading, Massachusetts, 1981.

[2] J.D. Ullman, Principles of Database Systems, Computer Science Press, Potomac, Maryland, 1980.

[3] F.W. Lancaster, Information Retrieval Systems--Characteristics Testing, Evaluation, Second Edition, John Wiley and Sons, New York, 1979.

[4] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw Hill Book Company, New York, 1983.

[5] P.H. Winston and R.H. Brown, Artificial Intelligence: An MIT Perspective, Vol. 1, MIT Press, Cambridge, Massachusetts, 1979.

[6] R.C. Schank and C. Riesbeck, Inside Computer Understanding: Five Programs plus Miniatures, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1981.

[7] J.C. Kolodner, Retrieval and Organizational Strategies in Conceptual Memory: A Computer Model, Ph.D. Dissertation, Computer Science Department, Yale University, New Haven, Connecticut, 1980.

[8] J. O'Connor, Retrieval of Answer Sentences and Answer Figures by Text Searching, Information Processing and Management, Vol. 11, No. 5/7, 1975, p. 155-164.

[9] J. O'Connor, Data Retrieval by Text Searching, Journal of Chemical Information and Computer Sciences, Vol. 17, 1977, p. 181-186.

[10] J. O'Connor, Answer Passage Retrieval by Text Searching, Journal of the ASIS, Vol. 31, No. 4, July 1980, p. 227-239.

[11] F.W. Lancaster and E.G. Fayen, Information Retrieval On-Line, Melville Publishing Company, Los Angeles, California, 1973.

[12] M. Casey, Packet Switched Data Networks: An International Review, Information Technology: Research and Development, Vol. 1, No. 3, July 1982, p. 217-244.

[13] M. Purser, The Euronet Diane Network for Information Retrieval, Information Technology: Research and Development, Vol. 1, No. 3, July 1982, p. 197-216.

[14] P.W. Williams, The Potential of the Microprocessor in Library and Information Work, Aslib Proceedings, Vol. 31, No. 4, April 1979, p. 202-209.

[15] A.D. Pratt, The Use of Microcomputers in Libraries, Journal of Library Automation, Vol. 13, No. 1, March 1980, p. 7-17.

[16] F.W. Lancaster, Toward Paperless Information Systems, Academic Press Inc., New York, 1978.

[17] C.T. Meadow, T.T. Hewett, and E.S. Aversa, A Computer Intermediary for Interactive Database Searching, I. Design, II. Evaluation, Journal of the ASIS, Vol. 33, No. 5, September 1982, p. 325-332, and Vol. 33, No. 6, November 1982, p. 357-364.

[18] R.S. Marcus and J.F. Reintjes, A Translating Computer Interface for End-User Operations of Heterogeneous Retrieval Systems, I. Design, II. Evaluation, Journal of the ASIS, Vol. 32, No. 4, July 1981, p. 287-317.

[19] V.V. Rao and P. Zunde, Some Approaches to Modeling Complex Information Systems, Information Processing and Management, Vol. 18, No. 3, 1982, p. 151-160.

[20] P. Zunde, Information Theory and Information Science, Information Processing and Management, Vol. 17, No. 6, 1981, p. 341-347.

[21] J. Tague, The Success-Breeds-Success Phenomenon and Bibliometric Processes, Journal of the ASIS, Vol. 32, No. 4, July 1981, p. 280-286.

[22] C.T. Yu, W.S. Luk, and M.K. Siu, On Models of Information Retrieval Processing, Information Systems, Vol. 4, No. 3, 1979, p. 205-218.

[23] M.E. Maron and J.L. Kuhns, On Relevance Probabilistic Indexing and Information Retrieval, Journal of the ACM, Vol. 7, 1960, p. 216-243.

[24] A. Bookstein and D.R. Swanson, Probabilistic Models for Automatic Indexing, Journal of the ASIS, Vol. 25, 1974, p. 312-319.

[25] S.P. Harter, A Probabilistic Approach to Automatic Keyword Indexing, Journal of the ASIS, Vol. 26, 1975, Part 1, p. 197-205, Part II, p. 280-289.

[26] J.M. Tague, A Bayesian Approach to Interactive Retrieval, Information Storage and Retrieval, Vol. 9, No. 3, March 1973, p. 129-142.

[27] W.S. Cooper and M.E. Maron, Foundations of Probabilistic and Utility Theoretic Indexing, Journal of the ACM, Vol. 25, No. 1, January 1978, p. 67-80.

[28] S.E. Robertson and K. Sparck Jones, Relevance Weighting of Search Terms, Journal of the ASIS, Vol. 27, 1976, p. 129-146.

[29] C.T. Yu and G. Salton, Precision Weighting--An Effective Automatic Indexing Method, Journal of the ACM, Vol. 23, 1976, p. 76-88.

[30] W.B. Croft and D.J. Harper, Using Probability Models of Document Retrieval without Relevance Information, Journal of Documentation, Vol. 35, No. 4, December 1979, p. 285-295.

[31] H. Wu and G. Salton, A Comparison of Search Term Weighting: Term Relevance vs. Inverse Document Frequency, ACM SIGIR Forum, Vol. 16, No. 1, Summer 1981, p. 30-39.

[32] C.J. van Rijsbergen, A Theoretical Basis for the Use of Cooccurrence Data in Information Retrieval, Journal of Documentation, Vol. 33, 1977, p. 106-119.

[33] D.J. Harper and C.J. van Rijsbergen, An Evaluation of Feedback in Document Retrieval Using Cooccurrence Data, Journal of Documentation, Vol. 34, 1978, p. 189-216.

[34] S.E. Robertson, C.J. van Rijsbergen, and M.F. Porter, Probabilistic Models of Indexing and Searching, in Information Retrieval Research, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen and P.W. Williams, editors, Butterworths, London, 1981, p. 35-56.

[35] C.T. Yu, C. Buckley, K. Lam, and G. Salton, A Generalized Term Dependence Model in Information Retrieval, Technical Report, Department of Computer Science, Cornell University, Ithaca, New York, 1983.

[36] C.T. Yu, W.S. Luk, and M.K. Siu, On Models of Information Retrieval, Information Systems, Vol. 4, No. 3, 1979, p. 205-218.

[37] H. Wu and G. Salton, The Estimation of Term Relevance Weights Using Relevance Feedback, Journal of Documentation, Vol. 37, No. 4, December 1981, p. 194-214.

[38] S.E. Robertson, M.E. Maron and W.S. Cooper, Probability of Relevance: A Unification of Two Competing Models for Document Retrieval, Information Technology: Research and Development, Vol. 1, No. 1, 1982, p. 1-21.

[39] W.S. Cooper and P. Huizinga, The Maximum Entropy Principle and its Application to the Design of Probabilistic Retrieval Systems, Information Technology: Research and Technology, Vol. 1, No. 2, April 1982, p. 99-112.

[40] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw Hill Book Company, New York, 1983, Chapter 3.

[41] G. Salton, A Blueprint for Automatic Indexing, ACM SIGIR Forum, Vol. 16, No. 2, Fall 1981, p. 22-38.

[42] A. Bookstein, Fuzzy Requests: An Approach to Weighted Boolean Searches, Journal of the ASIS, Vol. 31, No. 4, July 1980, p. 240-247.

[43] T. Radecki, Mathematical Model of Information Retrieval Based on a Concept of a Fuzzy Thesaurus, Information Processing and Management, Vol. 12, No. 5, 1976, p. 313-318.

[44] W.G. Waller and D.H. Kraft, A Mathematical Model for a Weighted Boolean Retrieval System, Information Processing and Management, Vol. 15, No. 5, 1979, p. 235-245.

[45] D.A. Buell and D.H. Kraft, Threshold Values and Boolean Retrieval Systems, Information Processing and Management, Vol. 17, No. 3, 1981, p. 127-136.

[46] G. Salton, E.A. Fox, and H. Wu, Extended Boolean Information Retrieval, Technical Report TR 82-511, Department of Computer Science, Cornell University, Ithaca, New York, August 1982.

[47] G. Salton, C. Buckley, and E.A. Fox, Automatic Query Formulations in Information Retrieval, Technical Report TR 82-524, Department of Computer Science, Cornell University, Ithaca, New York, October 1982.

[48] G. Salton, C. Buckley, E.A. Fox, and E. Voorhees, Autoamtic Relevance Feedback in Boolean Information Retrieval, Technical Report TR 83-539, Department of Computer Science, Cornell University, Ithaca, New York, January 1983.